Indeksowanie zawartości treściowej w bibliograficznych bazach danych

Wanda Klenczon
Biblioteka Narodowa

„Bibliograficzne bazy danych : kierunki rozwoju i możliwości współpracy”, Ogólnopolska konferencja naukowa z okazji 10-lecia bazy danych BazTech Bydgoszcz, 27-29 maja 2009 r.

Jakość oferowanych w bibliograficznych bazach danych punktów dostępu do treści dokumentów i efektywność wyszukiwania tych danych przez użytkowników zależy od wielu czynników, z których jako kluczowe należy wymienić:

- dobór odpowiedniego narzędzia do opracowania rzeczowego danego zbioru informacji, z uwzględnieniem:
  - jakości samego narzędzia
  - charakteru i rozmiar zbioru
  - kompetencji i oczekiwań użytkowników
  - możliwości i funkcjonalności systemu, w którym dane są gromadzone i udostępniane
- jakość opracowania rzeczowego:
  - trafna analiza dokumentu,
  - selekcja informacji
  - tłumaczenie na język informacyjno-wyszukiwawczy lub tekst adnotacji
  - utrzymanie spójności danych w obrębie całego zbioru.

Subject access points – punkty dostępu treściowego

- **języki informacyjno-wyszukiwawcze**
  języki sztuczne, który pełnią funkcję metainformacyjną (odwzorowują cechy informacji) i wyszukiwawczą (umożliwiają wyszukanie odpowiednich informacji)
  - języki haseł przedmiotowych, klasyfikacje, języki deskryptorowe i języki słów kluczowych

- **opisy rzeczowe wyrażone w języku naturalnym**,
  - abstrakty, adnotacje, streszczenia, słowa pochodzące z tekstu indeksowanych dokumentów

  pytanie o zasadność, użyteczność i ekonomikę tradycyjnego, manualnego opracowania rzeczowego

Języki haseł przedmiotowych – problemy i pytania o przyszłość

- pozorna naturalność słownictwa - użytkownik traktuje hasła jak wyrażenia języka naturalnego, nie zdając sobie sprawy ze sztuczności jhp
- sztuczna i zbyt skomplikowana gramatyka (reguły języków prekoordynowanych określają możliwość łączenia i sztywną kolejność poszczególnych elementów w haśle przedmiotowym)
- trudne do wyjaśnienia odstępstwa od reguł
- opracowanie w jhp jest trudne, czasochłonne i kosztowne

→ niezrozumienie haseł nie tylko przez użytkowników końcowych, ale i bibliotekarzy czy pracowników informacji naukowej nie będących specjalistami w zakresie opracowania rzeczowego

→ użyteczność haseł jest coraz częściej kwestionowana

Statystyki wykorzystania indeksów, analiza logów transakcyjnych, zachowań użytkowników i poziomu ich satysfakcji, wskazują na niewielkie wykorzystanie indeksów przedmiotowych (funkcja „browsing”) i dominującą pozycję wyszukiwania przez słowa lub kombinacje słów (funkcja „searching”)

- w bazach europejskich bibliografii narodowych – jedynie ok. 6-8% wyszukiwań haseł przedmiotowych, wyjątek (?) – baza katalogowa Biblioteki Narodowej: 20-25% wyszukiwań w indeksie przedmiotowym, jedynie 5-10% przez słowo/-a.
- badania logów transakcyjnych i zachowań informacyjnych użytkowników Oklahoma Univesity Library wykazały, że
  - jedynie 4,6% wyszukiwań realizowano za pośrednictwem indeksu przedmiotowego, 64,8% wyszukiwano przez słowa.
  - blisko 50% wyszukiwań przez hasło przedmiotowe było bezowocnych (0 trafień), 10% zapytań dało wynik ponad 500 pozycji;
  - użytkownicy, którzy nie są zadowoleni z efektów wyszukiwania nie próbują przeglądać indeksów ani nie korzystają z sugestii (linków czy odsyłaczy), ale zadają następne pytanie.

OPAC BN – statystyki wyszukiwań

Management Information on Public Catalog Searches

Indexes Used and Search Results From Tuesday 05 May 03:03AM, to Thursday 14 May 03:00AM		
	Number Done	**Percent Done**
RECORD NOS	764	0.20%
Other	1	0.00%
WORDS	13	0.00%
WORDS	16,361	4.30%
WORDS	11,861	3.12%
AUTHORS	98,204	25.79%
BARCODES	4	0.00%
CALL NOS	3,152	0.83%
SUBJECTS	74,226	19.49%
PUBLISHERS	702	0.18%
ISN'S	15,791	4.15%
Other	1	0.00%
NUKAT BIB NOS	41	0.01%
TITLES	134,065	35.21%
Other	25,580	6.72%
TOTAL	380,766	100.00%

Management Information on Public Catalog Searches

Search Results for User Keyed Searches - SUBJECTS From Tuesday 05 May 03:03AM, to Thursday 14 May 03:00AM		
	Number Done	**Percent Done**
Searches Retrieving 1 Record	34,060	45.89%
Searches Retrieving 2 to 8 Records	6,621	8.92%
Searches Retrieving 9 to 30 Records	4,934	6.65%
Searches Retrieving 31 to 99 Records	4,342	5.85%
Searches Retrieving 100 to 499 Records	5,702	7.68%
Searches Retrieving 500 to 4999 Records	3,773	5.08%
Searches Retrieving 5000 or More Records	665	0.90%
Total Searches Retrieving Records	60,097	80.96%
Total Records Retrieved/Average per Search	14,583,202	242
Searches With No Retrievals	14,129	19.04%
Total Searches	74,226	100.00%

Języki haseł przedmiotowych – dyskusja nad przyszłością Library of Congress Subject Headings

- ponad stuletnia tradycja – jhp projektowany dla katalogu kartkowego
- język prekoordynowany - zasady gramatyki pozycyjnej szczegółowo określa czterotomowy podręcznik „Subject Cataloging Manual”
- ponad 300 000 haseł wzorcowych, ok. 9 000 000 haseł przedmiotowych rozwiniętych w katalogu LC

Sugerowane kierunki rozwoju:

- aktualizacja i weryfikacja terminów i ich relacji,
- uproszczenie i uelastycznienie reguł gramatyki (np. rezygnacja z określników formalnych na rzecz zapisu ich w osobnym polu),
- adaptacja do wyszukiwania fasetowego (rozwój projektu FAST),
- udostępnienie zasobu w postaci SKOS (Simple Knowledge Organization System),
- zbadanie możliwości włączenia funkcji społecznego tagowania
- uzupełnienie opisów w LCSH o punkty dostępu pochodzące z innych języków informacyjno-wyszukiwawczych

Klasyfikacje
Języki deskryptorowe

Klasyfikacje o notacji numerycznej –.

- możliwość dostępu do informacji z pominięciem bariery językowej, która utrudnia międzynarodowe wykorzystanie języków informacyjno-wyszukiwawczych o notacji paranaturalnej
- ale: efektywnie można korzystać z tych charakterystyk wyszukiwawczych jedynie wtedy, gdy wszystkie symbole klasyfikacji zostaną opatrzone odpowiednikami słownymi.
- w wielu bazach bibliograficznych stosuje się własne, na ogół dość proste klasyfikacje służące zwłaszcza do organizacji zrębu bibliografii w wersji edycyjnej, ale także do ograniczania wyszukiwania czy uzupełniania innych metod wyszukiwana dokumentów

Języki deskryptorowe o notacji paranaturalnej

- indeksowanie współrzędne - prostsze niż budowa obwarowanych wieloma zasadami gramatyki pozycyjnej haseł przedmiotowych.
- poza alfabetycznym spisem deskryptorów, zawierają także część systematyczną, porządkującą terminy w strukturze hierarchicznej, co wydatnie pomaga zarówno twórcom tezaurusa rozbudowującym zasób o nowe terminy, indeksatorom poszukującym właściwych deskryptorów, jak i użytkownikom końcowym

Inne punkty dostępu treściowego

- **Abstrakty, streszczenia, adnotacje itp.**
  - dodatkowe dane o treści dokumentu, zwykle informacje o charakterze dokumentu, jego adresatach, głównych tezach, przyjętych metodach i uzyskanych wynikach. Zapisy te uzupełniają charakterystykę wyszukiwawczą, czasem ją zastępują.
  - sformułowane w języku naturalnym, w postaci tekstu uzupełniającego zasadniczy opis bibliograficzny lub będące częścią samego dokumentu, stanowią cenne źródło punktów dostępu do treści w bibliograficznych bazach danych, o ile są przeszukiwane przez system obsługujący bazę i odpowiednio indeksowane
  - ważnych informacji dostarczają także elementy wzbogacające opis bibliograficzny: dołączone spisy treści i zdjęcia okładek, recenzje czy komentarze, należące do nowej przestrzeni informacyjnej związanej z Siecią drugiej generacji.

Tagi, folksonomie - wartość dodana?

- nadawane przez odbiorców metadane, w których za pomocą swobodnych słów kluczowych lub wyrażeń wskazują treści, zawartość, kontekst i swój stosunek do oznaczanego dokumentu
- mogą sprawdzić się w wysoce specjalistycznych bazach dziedzinowych.
- szczególnie wartościowe i pomocne w wyszukiwaniu informacji mogą być tam, gdzie tradycyjne języki informacyjno-wyszukiwawcze są mało skuteczne, jeśli nie bezradne, np. w opisie dokumentów ikonograficznych

- wartość informacyjną tagów stawiają pod znakiem zapytania m.in.
  - brak rozróżnienia terminów wieloznacznych i homonimów,
  - występowanie synonimów,
  - błędy ortograficzne i niekonsekwencje w zapisie wyrażeń złożonych,
  - nieujednolicone formy gramatyczne (liczba pojedyncza i mnoga),
  - tagi emocjonalne, osobiste, symboliczne, nieetyczne, niecenzuralne.

Wyszukiwanie pełnotekstowe?
Indeksowanie automatyczne?

- Wyszukiwarki oferują już np.
  - selekcjonowanie najbardziej adekwatnych słów kluczowych,
  - obliczanie trafności zapytania i zwracanie najbardziej trafnych odpowiedzi na czołowych pozycjach wyników,
  - automatyczny przekład dokumentu
  - generowanie streszczeń

- Indeksowanie automatyczne jest realną przyszłością opracowania rzeczowego
  - w pierwszym rzędzie obejmie dokumenty elektroniczne dostępne sieciowo, których ilość wyklucza opracowanie manualne i dokumenty pełnotekstowe udostępniane w bibliotekach cyfrowych i bazach dziedzinowych

- Perspektywa całkowitego zastąpienia ręcznego indeksowania dokumentów bibliotecznych działaniem maszynowym jest w tej chwili trudna do wyobrażenia
  - względy techniczne (cały zasób dokumentów należy najpierw zdigitalizować),
  - ochrona prawem autorskim i prawami pokrewnymi uniemożliwia publiczny dostęp do większości nowszych publikacji (z wyjątkiem dostępu lokalnego lub odpłatnego)

Subject access points
w różnych typach bibliografii

- **Bibliografie narodowe**
  - zgodnie z zaleceniami międzynarodowymi oferują zwykle dwa rodzaje punktów dostępu treściowego:
    - jhp (LCSH, RAMEAU, RSWK, Nuovo Soggettario, JHP BN…) lub język deskryptorowy
    - jedna z klasyfikacji międzynarodowych (KDD, UKD, klasyfikacja Unesco)

- **Bibliografie regionalne**
  - komplementarne wobec bibliografii narodowej, opis rzeczowy często nawiązuje do standardów przyjętych w bibliotece narodowej danego kraju. Uniwersalny język informacyjno-wyszukiwawczy zwykle w takim przypadku jest modyfikowany zgodnie z lokalnymi potrzebami
  - dodatkowe punkty dostępu: własne klasyfikacje porządkujące układ bibliografii

- **Bibliografie dziedzinowe**
  - stosują specjalistyczne, dziedzinowe narzędzia indeksowania, zwykle tezaurusy i/lub klasyfikacje.
  - dodatkowe punkty dostępu: własne klasyfikacje porządkujące układ bibliografii

Subject access points
w różnych typach bibliografii

- **Bibliografie artykułów z czasopism -** zależnie od zakresu bibliografii i jej zasobności w dokumenty
  - bibliografie ogólne lub wielodziedzinowe - zwykle uniwersalny język informacyjno-wyszukiwawczy, modyfikowany pod kątem szczególnych potrzeb oraz własna kontrolowana lista słów kluczowych lub słowa z adnotacji i streszczeń
  - w wielu bazach oferujących dostęp do metadanych artykułów, a często i do ich treści, poprzestaje się na zastosowaniu dość ogólnej klasyfikacji, uzupełniając możliwości wyszukiwacze o słowa z tytułu, adnotacji, ewentualnie swobodne słowa kluczowe nadane przez autorów opisywanych tekstów
    - częstą, a zdecydowanie niekorzystną praktyką, jest ograniczenie wyszukiwania przez słowo do słów zawartych w tytułach

Tytuły publikacji jako źródło punktów dostępu do treści

TAK

- *Stosunek greckokatolickiego duchowieństwa do państwa polskiego w okresie II Rzeczypospolitej (1918-1939)*
- *Diagnoza sytuacji kobiet na rynku pracy w Lubuskiem*
- *Zarządzanie przez delegowanie uprawnień*
- *Ochrona małoletnich użytkowników mediów elektronicznych przez ustawodawstwo medialne Republiki Federalnej Niemiec*

???

- *Ciche bohaterki...*
- *Komentarz do dyrektywy Rady 2004/81/WE z 29.4.2004 r.*
- *Tryzub w labiryncie Minotaura*
- *Omnia aura mecum porto : pomyłka Benjamina a strategie artystycznej odnowy doświadczenia*

Funkcjonalny system wyszukiwania informacji o treści powinien:

- oferować funkcję
  - przeglądania indeksów (browsing),
  - wyszukiwania przez dowolne słowa (searching), z uwzględnieniem przeszukiwania całego opisu, jeśli zawiera on dodatkowe informacje o treści dokumentów
- umożliwiać obsługę
  - zapytań prostych (wyszukiwanie proste)
  - zapytań złożonych przy użyciu operatorów boolowskich (wyszukiwanie zaawansowane),
  - wyszukiwania według początków terminów lub zastosowaniem maskowania, wybór terminów z indeksu dla każdego pola,
- zapewnić link do wykorzystywanego zasobu słownictwa języka informacyjno-wyszukiwawczego,
- w miarę możliwości kontrolować słownictwo kartoteką wzorcową, z której w indeksie generowane są przynajmniej odsyłacze całkowite

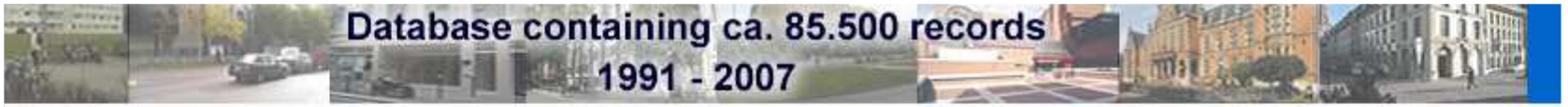
Nad czym warto się zastanowić…

- użycie więcej niż jednego narzędzia opisu rzeczowego oraz zaoferowanie zróżnicowanego poziomu indeksowania?
  - język haseł przedmiotowych/ język deskryptorowy i klasyfikację (np. wyszczególniające hasła przedmiotowe i uogólniające symbole klasyfikacji),
  - w uzasadnionych przypadkach dopuszczenie stosowania słownictwa niekontrolowanego

- uwzględnienie dodatkowych informacji?
  - punkty dostępu z adnotacji, abstraktów, streszczeń
  - dołączenie spisów treści dokumentów

Co jeszcze możemy zrobić?

- prezentować terminy wyszukiwawcze w postaci chmury tagów
- zaproponować użytkownikom współudział w opracowaniu rzeczowym poprzez zamieszczanie własnych tagów.
- opracować graficzne sposoby wyszukiwania, np. za pośrednictwem map czy wykresów, z których linki prowadzą do odpowiednich zasobów bibliograficznych lub tekstów
- zapewnić wyczerpującą, ale jasno sformułowaną informację o polityce i narzędziach indeksowania oraz możliwościach przeszukiwania zasobu (instrukcja wyszukiwawcza „just in time”, sugestia dalszych kroków w systemie)
- zaoferować limitowanie wyszukiwań i sortowanie wyników (fasetyzacja)
- uwzględnić system podpowiedzi (błędy ortograficzne!)
- wyjść w przestrzeń bibliograficzną poza obszar własnej bazy – np. przez linkowanie do/od zasobów zewnętrznych (Wikipedia, bazy faktograficzne, księgarnie internetowe)



EBSEES Subject headings as tag cloud

You can see the "weight" of used subject headings and click to perform a searching.

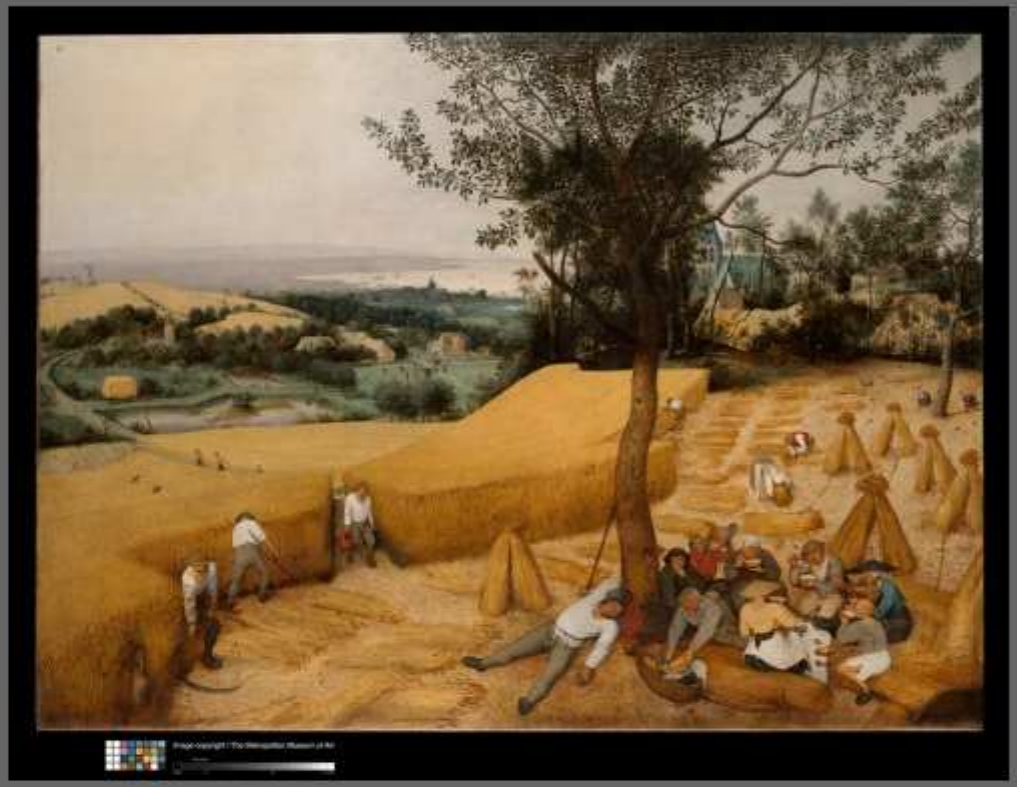
Academies of science Adaptations Aesthetics Agriculture **Agriculture, Fishing, Forests and Forestry** Air Transport Akaev, Askar **Albania** Albanian Emigration Aleksandr I, tsar of Russia Aleksievič, Svetlana Allied Occupation, 1945-1955 Americans Andrić, Ivo Annan, Kofi Anti-Semitism Aral Sea **Architecture Archives, Libraries, Museums, Information** Area Studies **Armenia** Armenian Christian Literature Arms control Art History **Arts, Painting, Photography** Assassination Auschwitz **Austria** **Austro-Hungarian Monarchy** Autobiography Azadovskij, Konstantin **Azerbaijan** Babits, Mihály Bakhtin, Mikhail Balázs, Béla **Balkans** Baltic **Baltic States** Banks and Banking Barkov, Ivan S. Bartoszewski, Władysław Bašmet, Jurij Bay, Marie-Pierre **Belarus** Belgium Belyi, Andrei Bérard-Zarzycka, Ewa Berger, Jakov Bessarabia **Bibliography, Reference Works, Area Studies** Bielecki, Czesław Biographies **Biography, Autobiography, Memoirs** Bismarck, Otto von Bloch, Jan Bocheński, Józef Maria Bohemia Bonner, Elena **Books, Publishing, Censorship** Bór-Komorowski, Tadeusz **Bosnia and Hercegovina** Boulgakov, see Bulgakov Brancusi, Constantin Brecht, Bertolt Brno Brosset, Marie-Félicité Bucharest Bujda, Jurij **Bulgaria** Bulgarian Burgenland Business Management Byron, George Gordon, lord Cambon, Fernand Čapek, Karel Carpathian Mountains Caspian Sea Region Catholics **Caucasus** Ceauşescu, Nicolae Center-Periphery Relations **Central Asia** **Central Europe** Chagall, Marc Chechnia Child Welfare Chruschtschow, see Hruščev Churches Čikatilo, Andrej R. **Cinema, Film, Video** **CIS** **Cities and Towns, Urban Life, Rural Life, Peasantry** Citizenship Civil War, 1992-1997 Clément, Olivier Collectivization Commercial law Communist Party of the Soviet Union Concentration Camps Constantinescu, Emil **Constitution, Public Administration, Elections** Cooking Cost and Standard of Living Craft industry Criticism and Interpretation **Croatia** Cuba Cultural History **Cultural Relations** **Culture** Cvetaeva, Marina I. **Czech Republic** Czechoslovakia Dagestan Dance, Ballet Darvasi, László Dead Souls Defence industry Deml, Jakub **Demography, Population, Migration** Denmark Description and Travel Diaspora Outside Eastern Europe Dimitrij Ivanovič, prince of Russia Disarmament **Dissent, Opposition** Djagilev, Sergej P. Dobosz, Andrzej Doctor Zhivago Dominiak, Zbigniew **Dostoevskij, Fedor M.** Dostoevsky, Fyodor Drugs Dudaev, Džohar Durova, Nadežda A. Early printed books **Eastern Europe** **East-West Trade** Ebner-Eschenbach, Marie von **Ecology, Environment, Nuclear Disaster, Pollution** **Economic Relations** Economic Stabilization **Economy** **Education, Scholarship, Schools, Universities** Educational Exchanges Eisenhower, Dwight D. Elderly Elizabeth, empress of Austria-Hungary **Emigration** Emigré Literature, Samizdat Energy **Energy, Industry, Building, Handicraft** Entrepreneurs Erdman, Nikolaj R. Esenin, Sergej A. **Estonia** Estonian Ethnography Europe **European Community** European Union **European Union Eastern**

LOG OFF

THE ART MUSEUM SOCIAL TAGGING PROJECT

SET 1

581 REMAINING



Pieter Bruegel the Elder, Netherlandish, active by 1551, died 1569
The Harvesters, 1565
Oil on wood; Overall, including added strips at top, bottom, and right, 46 7/8 x 63 3/4 in. (119 x 162 cm); original painted surface 45 7/8 x 62 7/8 in. (116.5 x 159.5 cm)
Rogers Fund, 1919, 19.164
The Metropolitan Museum of Art

Tags for this work:

Dutch, Seasons last hay, aerial perspective, agriculture, amber, autumn, black birds, country, crop, cut, details, field, folk art, genre, grain, harvest, hay, hazy, landscape, lunch, medieval, peasant, peasants, perspective, pesants, picnic, rest, sheaves, sickle, siesta, sleep, sleepy, snack, snooze, summer, thresh, tree, trees, warm field, work

Start Here

ADD

NEXT

K TO SETS

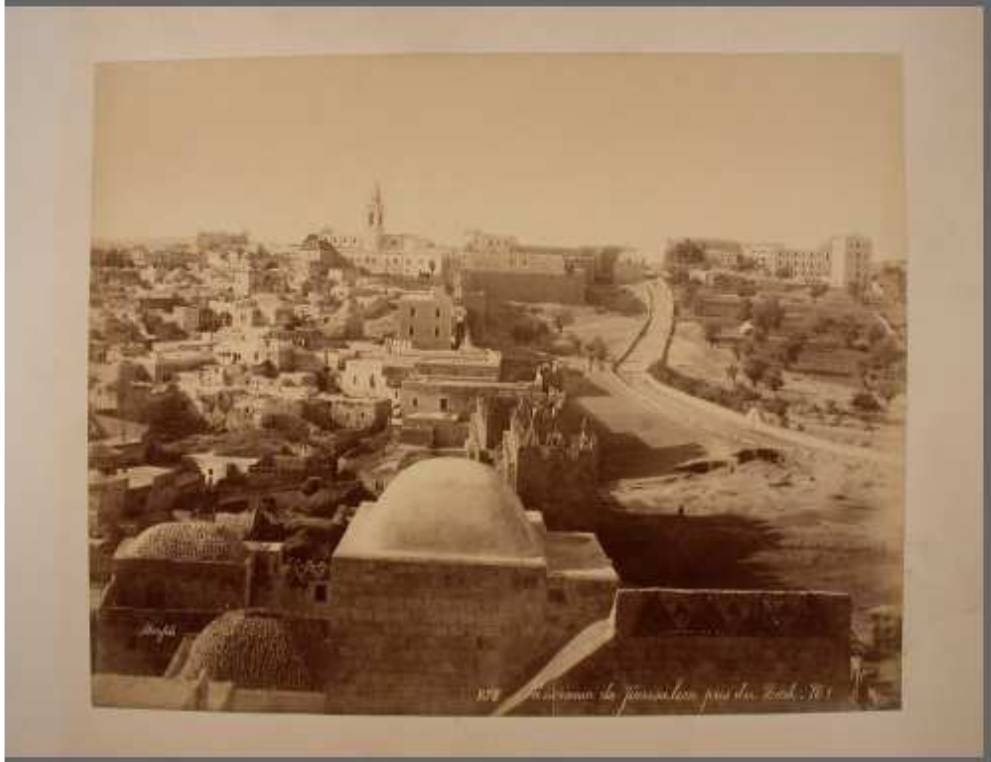
OFF

steve

THE ART MUSEUM SOCIAL TAGGING PROJECT

ET 2

214 REMAINING



elix Bonfils
hotograph
68.162
kirball

Tags for this work:

19th century, Arab city, French, Jerusalem, Religious sites, Sepia-toned photo, albumen, architecture, are all people dead?, arid, bonfils, brick and mortar architecture, caption, chapel, church, cities, city, city layout, city view, cityscape, desert, dome, domed, domed building, early photography, european, fortification, grecain, holy land, horizon, landscape, living quaters, long road, old, panorama, panoramic, photograph, photography, postcard, road, rooftop

Start Here

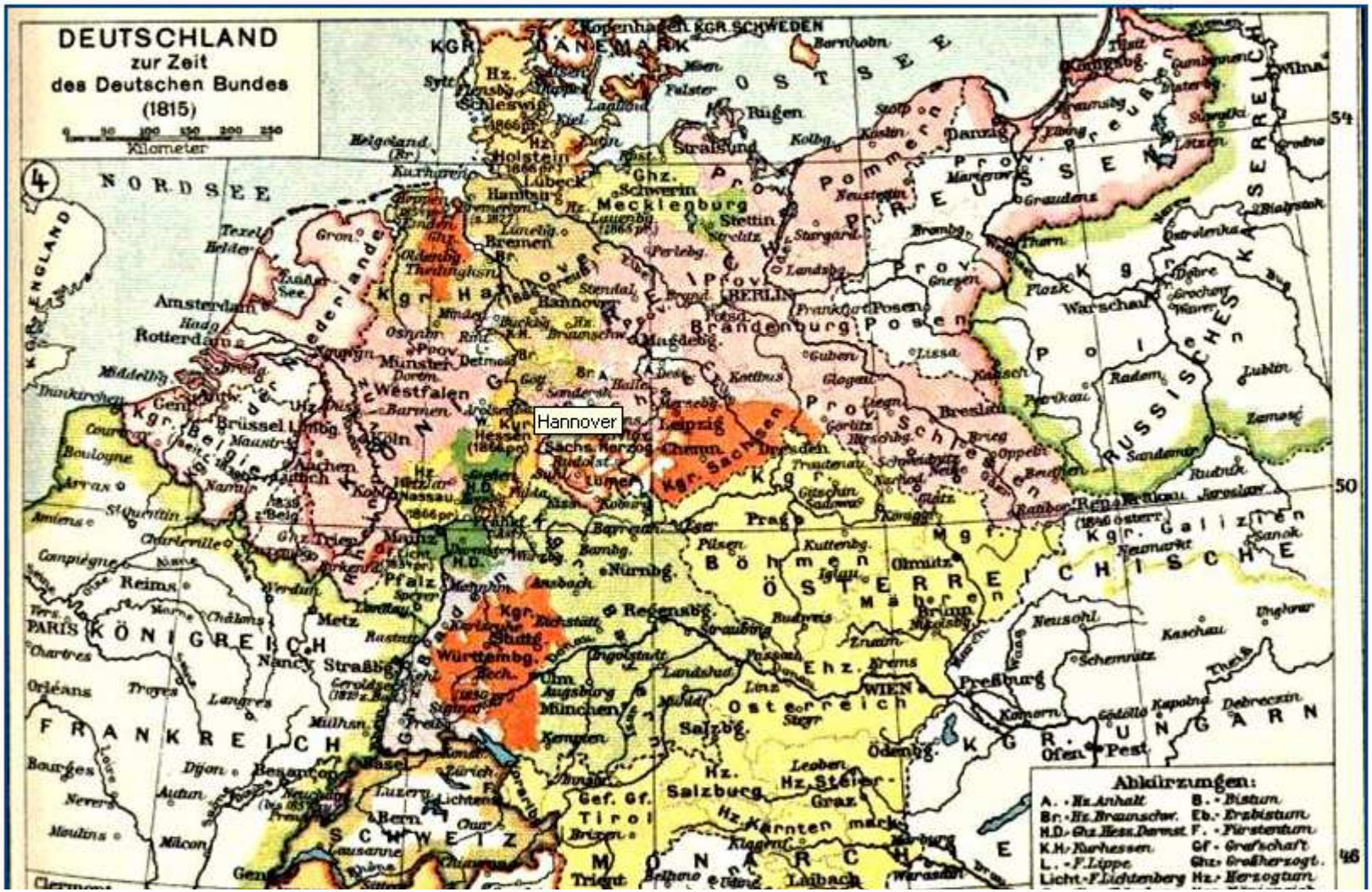
ADD

NEXT

Projekt Historischer Roman

Datenbank

Durch das Anklicken der Karte erhalten Sie die Autorinnen und Autoren der jeweiligen Region. Die Aufnahme der Geburtsorte erfolgte nach den politischen G

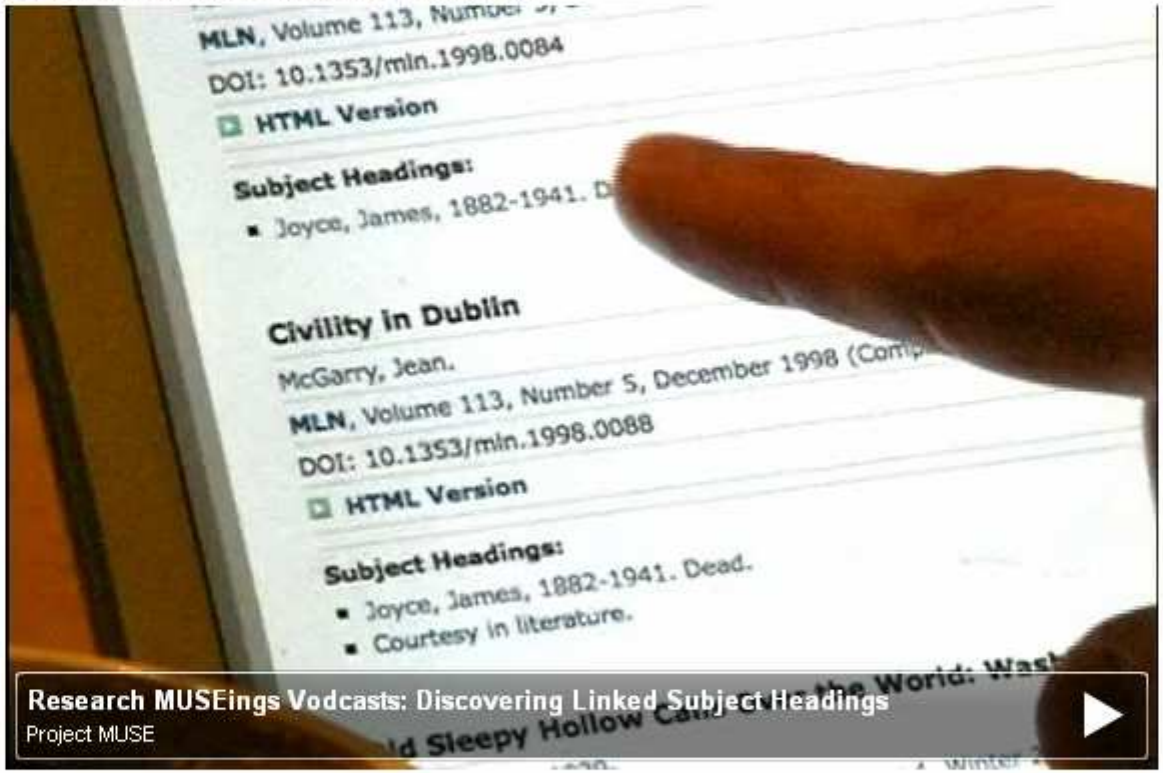


Tools & Resources

- Overview of Tools & Resources
- RSS Feeds
- Email Alerts
- Search Plug-in
- Bookmarking and Sharing
- Training & Usage Guides

Research MUSEings Vodcasts

Discovering Linked Subject Headings



Video podcasts produced by MUSE to help enhance your research on the site.

Searches with No Direct Hits - SUBJECTS
ok. 40% to błędy pisowni

- afryka subsacharyjska
- afryuka g
- agrsja
- akcyza procedory rozliczen
- aktywizacjca
- aktywizajca
- akty stenu cywilnego
- alzhaimer
- anatamia
- angeologia
- architekura
- badaniafaz
- budybki biblioteczne

Czy chodziło Ci o: ……?

Zamiast podsumowania

- Użytkownik ma prawo oczekiwać wyczerpującej informacji o treści dokumentów rejestrowanych w bibliografiach
- Dobór właściwego narzędzia/narzędzi opisu rzeczowego jest jedną z kluczowych decyzji podejmowanych przez twórców bibliograficznych baz danych
- Zadaniem indeksatora jest dostarczenie wiarygodnej, wysokiej jakości, łatwo dostępnej informacji o treści dokumentów
- Nowe technologie nie stanowią zagrożenia, ale wyzwanie i szansę
- Stała i dobra współpraca z informatykami, twórcami i administratorami systemów, jest dziś *conditio sine qua non* rozwoju bibliograficznych baz danych i efektywnego wyszukiwania informacji o treści

ciągle jeszcze …
„Subject analisys is
a core function of cataloging”

(„On the record…” LC 2008)

Dziękuję za uwagę

Wanda Klenczon
Biblioteka Narodowa
w.klenczon@bn.org.pl